DOCUMENT RESUME

ED 280 898 TM 870 231

AUTHOR Weare, Jane; And Others
TITLE Interrater Reliability: A Selected and Annotated Bibliography of Articles Concerning Interrater Reliability.
PUB DATE [87]
NOTE 17p.
PUB TYPE Reference Materials - Bibliographies (131)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Annotated Bibliographies; Cloze Procedure; Correlation; Essay Tests; *Interrater Reliability; *Mathematical Models; Rating Scales; Test Reliability; Test Validity

ABSTRACT

This annotated bibliography was developed upon noting a deficiency of information in the literature regarding the training of raters for establishing agreement. The ERIC descriptor, "Interrater Reliability", was used to locate journal articles. Some of the 33 resulting articles focus on mathematical concepts and present formulas for computing interrater reliability, while other articles are study reports and literature reviews. (LMO)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED280898

Interrater Reliability:
A selected and annotated bibliography of articles concerning interrater reliability

Jane Weare
Julia Moore
Frederick Woodall

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. Weare

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

TM 870 231

Upon noting a deficiency of information in the literature regarding the training of raters for establishing agreement this review was begun. The ERIC descriptor, interrater reliability, was used to locate journal articles. In this bibliography, some articles focus on mathematical concepts and present formulas for computing interrater reliability while other articles are study reports and literature reviews. For anyone interested in interrater reliability, more efficient use of time and resources can result from use of this annotation.

J. Weare

J. Moore

F. Woodall

Interrater Reliability: A selected and annotated bibliography of articles concerning interrater reliability

1. Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. Psychological Reports, 19, 3-11.

Bartko explains a procedure for estimating the reliability of a set of ratings using analysis of variance. Using the intraclass correlation coefficient (ICC) definition and an estimate of R in terms of mean squares, it is demonstrated that the ratio of variances is equivalent to and can be interpreted as a correlation coefficient. The one-way classification presents tables which describe data organization. Appropriate formulas are given with identified data sources for substitutions. Bartko also applies the two-way classification and provides estimates for the variance components. Emphasis is placed on the handling (or mishandling) of data and definitions and on the assumptions (or violations) regarding interaction when applying the two-way model. Illustrative examples using formulas and tables are also provided for this model.

2. Bartko, J. J. (1976). On various intraclass correlation reliability coefficients. Psychological Bulletin, 83, 762-765.

Bartko identifies and argues for intraclass correlation reliability coefficients (ICC) obtained from analysis of variance. The analysis of variance ICC as a reliability measure avoids what Bartko describes as serious deficiencies often found in reliability measures. The effects of biased data on the reliability measures obtained from Winer's anchor point model, the K-R 20, the Spearman Brown ICC, and Fleiss' ICC are compared with Bartko's ANOVA-based ICC. Noting that small within subjects variance should yield a high intraclass correlation

reliability coefficient, examples were used to illustrate the inconsistency of results across reliability measures. When the various procedures were applied to dichotomous data, results again were not stable. There is pointed discussion regarding the use and misuse of techniques applied to dichotomous data.

3. Blok, H. (1985). Estimating the reliability, validity, and invalidity of essay ratings. Journal of Educational Measurement, 22, 41-52.

Blok discussed study results concerning the relationship between multiple ratings and true scores. Data for the reliability of the essay marking were provided by 16 elementary teachers who independently rated 105 essays on a 10 point scale on two different occasions. Rater equivalence was explored using four models: parallel, tau-equivalent, Wiley and Wiley, and congeneric. Using parameters of the best fitting model, coefficients of reliability (intrarater), validity, and invalidity were computed. Suggestions for further research with cautions and limitations are given.

4. Cicchetti, D. V., Heavens, R., Didrikson, J., & Showalter, D. (1984). A computer program for assessing the reliability of nominal scales using varying sets of multiple raters. Educational and Psychological Measurement, 44, 671-675.

This article describes a computer program which from nominal data computes kappa, proportion of interrater agreement (PO), and proportion of interrater agreement expected by chance alone (PC). For each kappa value, statistical significance and clinical or substantive significance are also given. The approach is referenced, and program features with limitations are explained. This program is free from the authors and the address is given.

5. Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 37-46.

Cohen presents kappa as a coefficient of agreement and develops the argument for interpreting it as a measure of agreement (two raters). Sampling characteristics including significance are discussed. A practical explanation is given using an example. (See Fleiss, Cohen, & Everitt, 1969: Standard error formulas are incorrect in the above (Cohen, 1960) article.)

6. Cohen, J. (1968). Weighted kappa; Nominal scale agreement with provision for scaled disagreement and partial credit. Psychological Bulletin, 70, 213-220.

Kappa as a coefficient of agreement (2 raters) has been modified to consider when disagreement varies in seriousness. The weighted kappa is developed and presented with all arguments for its being "fully" chance corrected like the statistic kappa. An illustrative example is used and explained. (See Fleiss, Cohen, & Everitt, 1969: Standard error formulas are incorrect in the Cohen (1968) article.)

7. Collis, G. M. (1985). Kappa, measure of marginal symmetry and intraclass correlations. Educational and Psychological Measurement, 45, 55-62.

Collis explores Cohen's kappa coefficient. Measuring marginal symmetry along with kappa is suggested. Marginal symmetry should be examined to clarify disagreement. Intraclass correlation coefficients (ICC) for dichotomous data are presented as the complement of M, a measure of marginal symmetry.

8. Cooper, H. (1985). The effects of familiarity, gender, and institutional prestige on evaluative judgments of convention program proposals. Journal of Research and Development in Education, 18, 25-28.

This study focused on the reliability of convention proposal reviews. The study also addressed the influence of reviewer familiarity with the proposal content and of the gender and institutional affiliation of the first author. Findings indicated that the reliability was essentially like that reported for journal manuscripts. Reviewer familarity with content and/or author(s) had little impact on the evaluation. Familiarity was a factor in judging the appropriateness of the paper for a special interest group's program. Reviewer gender was not a factor in evaluation. Author gender was a factor since it was noted that female authors received lower evaluations regardless of the gender of the reviewer. An author's institutional ranking was of interest. Papers from unranked institutions were rated as having better designs and more valid links between results and conclusions. Cooper suggests caution when interpreting these findings.

9. DiSanti, R. J., & Sullivan, V. G. (1985). Reliability of single-rater judgments of semantic and syntactic classifications of cloze test responses. Journal of Research and Development in Education, 18, 49-53.

The acceptability of a response on a cloze test is subjective when the quality of a reader's response is evaluated. DeSanti and Sullivan investigated single rater consistency using qualitative cloze-based assessments of reading comprehension. There were four passages for each of 5 grade levels developed using established techniques. for a cloze test of reading comprehension. Traditional Comprehension,

Logical Language Usage, and Grammatically Unacceptable were the four interpretive categories analyzed. Seven raters used the same scoring directions to rate reader responses . Reliability was computed using mean squares. The overall reliability of ratings of reading comprehension from the Cloze Reading Inventory was not lowered by the potential subjectivity of ratings.

10. Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. Psychological Bulletin , 76, 378-372.

Fleiss develops a generalization of the statistic kappa. Kappa, when introduced by Cohen (1960) was a nominal scale measure of agreement between 2 raters and disagreement was assumed to be equally serious. Each rater judged all cases. Weighted kappa (Cohen, 1968) was developed for use when the relative disagreement between 2 raters judging all cases can be specified. This article presents the generalization of unweighted kappa as a measure of agreement among any constant number of independent raters. A procedure for identifying overall agreement corrected for the amount expected by chance is presented. Agreement among raters assigning a subject to a particular category is discussed with regard to its happening by chance. (See Fleiss, Nee, & Landis, 1979: Standard error formulas in Fleiss (1971) are incorrect.)

11. Fleiss, J. L., Cohen, J., & Everitt, B. S. (1969). Large sample standard errors of kappa and weighted kappa. Psychological Bulletin, 72, 323-327.

This article identifies the incorrectness of expressions for standard error found in earlier articles. Usable valid formulas are given.

An example is provided. (Also see Fleiss, Nee, & Landis, 1979)

12. Fleiss, J. L., Nee, J. C. M., & Landis, J. R. (1979). Large sample variance of kappa in the case of different sets of raters. Psychological Bulletin, 86, 974-977.

Fleiss, Nee, and Landis have presented new formulas for the large sample variance of the statistic kappa. They state that "in this article, formulas for the standard error of kappa in the case of different sets of equal numbers of raters that are valid when the number of subjects is large and the null hypothesis is true are derived. The results of some Monte Carlo simulations confirm that the formulas are correct" (p. 974). It is also shown that kappa is equivalent to the intraclass correlation coefficient except for terms which go to zero as the number of subjects increases.

13. Floyd, F. J., & Markham, H. J. (1983). Observational biases in spouse observation: Toward a cognitive/behavioral model of marriage. Journal of Consulting and Clinical Psychology, 51, 450-457.

This article explores the consistency of ratings from married couples and from observers. Spouse's ratings of partner were not consistent with observer's ratings of partners' behaviors. Observer ratings, however, were consistent with spouse's rating of own behavior. Explanation of method and procedure is given. Interpretation of results acknowledges limitations.

14. Fuchs, D., Fuchs, L., Dailey, A. M., & Power, M. H. (1985). The effects of examiners' personal familiarity and professional experience on handicapped children's test performance. The Journal of Education Research, 78, 141-146.

Regardless of the examiner's experience, handicapped preschoolers had stronger test performance with a personally familiar examiner. This paper has implications for assessing handicapped youngsters in terms of current testing practices.

15. Gobout, P., & Schutz, R. W. (1983). Generalizability of ratings of motor performances with references to various observational designs. Research Quarterly for Exercise and Sport, 54, 20-27.

Several generalizability coefficients are examined for appropriateness using observational data where each subject is rated by the same observers over a number of trials. Assumptions regarding error sources are noted.

16. Green, K. (1985). The paired comparison method in educational research. Educational and Psychological Measurement, 45, 567-575.

This study examines intransitivity in terms of the appropriateness of paired comparison judgements.

17. Halpin, G., Sigmon, G., & Halpin, G. (1983). Minimum competency standards set by three divergent groups of raters using three judgmental procedures: Implications for validity. Educational and Psychological Measurement, 43, 185-196.

Previous studies indicate that standards set varied with the method used which raises questions regarding the validity of the resulting decisions. This study centered on selecting a procedure (Angoff, Ebel, or Nedelsky methods) for setting standards that would allow valid decisions to be made. Standards set by raters at different competency levels as well as the effects of the interaction between rater group and procedure on standard set were explored.

18. Henk, W. A., & Selders, M. L. (1984). A test of synonymic scoring of cloze passages. The Reading Teacher, 38, 282-287.

Cloze tests have been used as indicators of reading comprehension levels. Scoring requires that an exact word be supplied in order to be considered correct. This study investigated rater agreement on accepting synonyms when scoring a cloze test. Results indicate scoring for synonyms is too varied to be practical.

19. Irvine, J. J. (1983). The accuracy of pre-service teacher's assessments of their classroom behaviors. Journal of Research and Development in Education, 17, 25-31.

Irvine investigated the relationship between the self-assessment of preservice teachers and supervising teachers' assessment of the preservice teachers classroom performance. The preservice teachers were trained in self-assessment and collegial relationships. Results indicate moderate agreement in the ratings.

20. Mayfield, K. L., Forman, S. G., & Nagle, R. J. (1984). Reliability of the AAMD Adaptive Behavior Scale-Public School Version. The Journal of School Psychology, 22, 53-61.

Using the AAMD Adaptive Behavior Scale - Public School Version to observe special education students, this study reports on the reliability of ratings and the stability of the instrument. Interrater reliability was established using 4 types of observers: Special education teachers, regular classroom teachers, parents, and an independent observer (graduate student). Observers provided 2 ratings with a 2 week interval between ratings. Ratings were stable over time, but there were differences between raters. Rater bias entered into the interpretation of the interrater reliability. The vested interest of parents and differences in the observation expertise of the classroom teachers and the special education teachers were noted. Also the child's behavior may actually change i.e., be situation specific, depending upon the variations in environments.

21. Nevo, B. (1985). Face validity revisited. Journal of Educational Measurement, 22, 287-293.

Nevo presents a review of the literature and develops an argument for face validity as a measured test characteristic.

22. O'Hara, M. W., & Rehm, L. P. (1983). Hamilton Rating Scale for Depression: Reliability and validity of judgments of novice raters. Journal of Consulting and Clinical Psychology, 51, 318-319.

The Hamilton Rating Scale for Depression provides an estimate of the severity of clinical depression. When used in research, at least

2 clinicians serve as raters, and interrater agreement must be demonstrated each time. This study examined the feasibility of training a novice rater to serve as a second rater with a clinician. The intraclass correlation coefficient (see Bartko, 1976) was used as the measure of interrater agreement and to establish the validity of judgments of novice raters. The training method is explained. Results indicated that novice raters with minimal training can make reliable and valid judgments of depression severity using this scale.

23. Orwin, R. G., & Corday, D. S. (1985). Effects of deficient reporting on meta-analysis: A conceptual framework and reanalysis. Psychological Bulletin, 97, 134-147.

This article examines the effects of reporting quality on meta-analytic study results. Discussion of establishing interrater reliability in terms of selection of variables for meta-analysis is provided. Coder uncertainty and coder error contribute to decreased reliability. Coder uncertainty is directly linked to decreased confidence. Results from the recoding and reanalyzing of 25 reports sampled from an earlier meta-analysis study demonstrated the influence of deficient reporting on meta-analytic data.

24. Paunonen, S. V., & Jackson, D. N. (1985). Idiographic measurement strategies for personality and prediction: Some unredeemed promissory notes. Psychological Review, 92, 486-511.

This article deals with predicting behavior and behavior consistency across situations. Interrater agreement between self and peer ratings of the levels of behaviors exhibited by subjects was used in the data base.

25. Payne, B. D. (1984). Interrelationships among college supervisor, supervising teacher, and elementary pupil ratings of student teaching performance. Educational and Psychological Measurement, 44, 1037-1043.

Pupil ratings of student teachers were compared with ratings provided by supervising teachers and college supervisors. Results indicated pupil ratings were not valid evaluators of student teacher performance. College supervisor's ratings and the supervising teacher's ratings were in "modest" agreement, but far different from pupil ratings.

26. Rae, G. (1984). On measuring agreement among several judges on the presence or absence of a trait. Educational and Psychological Measurement, 44, 247-253.

Rae discusses the special case of multiple raters judging the presence or absence of a trait (dichotomous response) in subjects. Four measures of agreement are shown to be mathematically equivalent in this special case. An example is provided.

27. Scott, M. M., & Hatfield, J. G. (1985). Problems of analyst and observer agreement in naturalistic narrative data. Journal of Educational Measurement. 22, 207-218.

This article discusses the problems associated with computing analyst and observer agreement using data from observations within natural settings. A review of methods for calculating agreement is presented with advantages and disadvantages of possible solutions.

28. Smith, P. K., & Vollstedt, R. (1985). On defining play: An empirical study of the relationship between play and various play criteria. Child Development, 56, 1042-1050.

Rater agreement on five criteria associated with play behavior was used to judge video tapes of nursery school children ages 3 and 4 years. Ratings provided the data base for identifying the criteria associated with play. Judgments regarding the presence or absence of play behavior were also analyzed.

29. Turner, S. M., Beidel, D., C., Hersen, M., & Bellack, A. S. (1984). Effects of race on ratings of social skill. Journal of Consulting and Clinical Psychology, 52, 474-475.

Black and white judges were trained according to the same procedure, and they rated subjects on social skills using the same scoring criteria. Analysis of their ratings indicated reliability within each racial group and differences between the racial groups. There were also interactions (race of judge x race of subject) which have implications for researchers to consider.

30. Weinrott, M. R., & Jones, R. R. (1984). Overt versus covert assessment of observer. Child Development, 55, 1125-1137.

This was an investigation of sustained reliability following training. Raters were trained to use one of four observation systems (all complex). Interrater reliability (ICC) was established during training. The raters were later monitored to determine whether reliability was maintained at the established level. Across stystems there was a

pattern of consistently lower reliability during the covert assessment. Pairing observers (overt assessment) indicated that observers coding with a partner increased reliability. An observer coding without a partner appears to be less attentive.

31. Whitehurst, G. J. (1984). Interrater agreement for journal manuscript reviews. American Psychologist, 39 22-28.

Whitehurst has looked at the manuscript review process for scholarly journals and offered Finn's r as an appropriate approach to computing agreement. He argues that poor manuscripts and high rejection rates reduce variance which cause an intraclass correlation to be low and misleading.

32. Wilson, D. R., & Griswold, M. L. (1985). The effects of method and comprehensiveness of training on the reliability and validity of ratings of counselor empathy. Measurement and Evaluation in Counseling and Development, 18, 3-11.

The results of a comparison of training methods is reported in this article. Error training and accuracy training were combined with comprehensiveness of training and with the underlying distribution for the criterion task. The accuracy training method increased the validity of observations. There were no significant interactions.

33. Wilson, F. R., Griswold, M. L., & Sunderland, E. (1984). Observer usage in counseling research: An analysis of reporting adequacy. Measurement and Evaluation in Guidance, 17, 40-43.

The data for this study come from a review of the reports of empirical studies in 5 counseling and psychology journals during a 5 year period. The use of the observer, observer training, and observer reliability and validity were categories of interest. Reports of type and length of training were insufficient to nonexistent in a majority of articles. Thus, insufficient information is provided for replication of the reviewed studies.